MEDICAL IMAGE CAPTIONING

USING TRANSFORMERS

# A MINI PROJECT-II REPORT

***Submitted by***

# BHARATH V S (71812111011)

# MARIAPPAN V (71812111027)

**MUTHUSELVAN B (71812111033)**

***in partial fulfillment for the award of the degree of***

# BACHELOR OF TECHNOLOGY

***in***

# ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

**SRI RAMAKRISHNA ENGINEERING COLLEGE**

**[**Educational Service: SNR Sons Charitable Trust] [Autonomous Institution, Reaccredited by NAAC with ‘A+’ Grade]­

[Approved by AICTE and Permanently Affiliated to Anna University, Chennai] [ISO 9001:2015 Certified and All Eligible Programmers Accredited by NBA] Vattamalaipalayam, N.G.G.O, Colony Post,

Coimbatore-641022

# ANNA UNIVERSITY CHENNAI 600 025

**MAY 2024**

**BONAFIDE CERTIFICATE**

**20AD277 – Mini Project - II**

Certified that this mini project Report **“MEDICAL IMAGE CAPTIONING USING TRANSFORMERS”** is the bonafide work of **“BHARATH.V.S (71812111011), MARIAPPAN.V (71812111027), MUTHUSELVAN. B (71812111033)” who** carried out theproject under my supervision.

|  |  |
| --- | --- |
| **SIGNATURE** | **SIGNATURE** |
| Dr. V. Karpagam | Mrs.P.V.Kavitha |
| **HEAD OF THE DEPARTMENT** | **SUPERVISOR** |
| Professor, | Assistant Professor (Sl.Gr) |
| Department of Artificial Intelligence and  Data Science | Department of Artificial Intelligence and  Data Science |
| Sri Ramakrishna Engineering College | Sri Ramakrishna Engineering College |
| Vattamalaipalayam | Vattamalaipalayam |
| Coimbatore-22 | Coimbatore-22 |

**Submitted for the Mini Project viva-voce examination held on**

**INTERNAL EXAMINER EXTERNAL EXAMINER**

# ACKNOWLEDGEMENT

We thank the almighty for his blessing on us to complete this project work successfully.

With profound sense of gratitude, we thank the management, Managing Trustee, **Thiru D**.**Lakshminarayanaswamy B. Tech., MBA**, and Joint Managing Trustee **Thiru R. Sundar**, for having provided us the necessary infrastructure required for completion of our project.

With profound sense of gratitude, we sincerely thank the **Head of the Institution**, **Dr. N.R. Alamelu M.E., Ph.D.**, for her kind patronage, which helped in pursuing the project successfully.

With immense pleasure, we express our hearty thanks to the **Head of the department**, **Dr. V. Karpagam M.E., Ph.D.**, Department of Artificial Intelligence and Data Science for her encouragement towards the completion of this project.

We thank our Project Coordinator **Mrs. P.V. Kavitha M.E.**, **Assistant Professor (Sl.G),** Department of Artificial Intelligence and Data Science for her encouragement and towards the completion of this project.

We also thank our Mini project guide **Mrs. P.V.Kavitha M.E., Assistant Professor (Sl.G)**, Department of Artificial Intelligence and Data Science for her guidance towards the completion of this Project.

We convey our thanks to all the teaching and non-teaching staff members of our department who rendered their co-operation by all means for completion of this project.

# ABSTRACT

This project pioneers a novel approach in medical imaging by merging cutting-edge technologies: Vision Transformer (ViT) and Generative Pre-trained Transformer 2 (GPT-2). By leveraging the Radiology Comments (ROCO) dataset, which pairs radiological images with expert annotations, the goal is to automatically generate accurate and contextually relevant descriptions for these images. This method involves fine-tuning a pre-trained ViT model on the ROCO dataset to extract meaningful visual features, which are then fed into a GPT-2 model to produce detailed captions. The proposed approach has been evaluated using standard metrics like BLEU, ROUGE, and METEOR scores, comparing it against baseline methods and conducting qualitative analysis on the generated captions. The experiments affirm the effectiveness of the ViT-GPT-2 fusion in generating informative and coherent descriptions for medical images. This underscores its potential in clinical settings to aid radiologists and healthcare professionals in image interpretation and decision-making

# TABLE OF CONTENTS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CHAPTER NO** | | **TITLE PAGE NO** | | |
| **ABSTRACT** 4  **LIST OF FIGURES** 7  **LIST OF TABLES** 8  **LIST OF ABBREVIATIONS** 9  **1 INTRODUCTION** 10 | | |  |
|  | | 1.1 ViT-GPT2 Model 10 | | | |  |
| **2** | | **LITERATURE SURVEY** 11 | | | |  |
| **3** | | **SYSTEM ANALYSIS** 16 | | | |  |
|  | | 3.1 EXISTING SYSTEM 16 | | | |  |
|  | | 3.2 PROPOSED SYSTEM 17 | | | |  |
| **4** | | **PROBLEM DEFINITION** 18 | | | |  |
|  | | 4.1 PROBLEM DEFINITION 18 | | | |
|  | | 4.2 OBJECTIVES 18 | | | |
|  | | 4.3 OVERVIEW OF THE PROJECT 19 | | | |
|  | | 4.4 BLOCK DIAGRAM 21 | | | |
|  | | 4.5 MODULE DESCRIPTION 22  4.5.1 Data Collection 23  4.5.2 Data Preprocessing And  Feature Extraction 23  4.5.3 Tokenization 23  4.5.4 Model Development And  Training 24  4.5.5 Validation 25 | | | |
| **5** | | **PERFORMANCE EVALUATION** 27 | | | |  |
|  | 5.1 RESULT 28  5.2 ACCURACY 32    **6** **CONCLUSION** 33  6.1 CONCLUSION  6.2 FUTURE SCOPE 34  **7 REFERENCES** 35  **8 APPENDIX** 36 | | | | |  |

**LIST OF FIGURES**

**FIGURE NO. TITLE PAGE NO**

4.1 Flow Diagram 21

4.2 ViT Architechture 21

4.3 GPT2-Architechture 22

4.4 Training Validation Loss 26

4.5 Rouge Validation Metric 26

5.1 BLEU Scores 27

5.2 N-Gram Analysis 29

5.3 Actual Captions 30

5.4 Predicted Captions 31

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **TABLE NO** | **TITLE** | **PAGE**  **NO** |
| 4.1 | MODEL TRAINING | 25 | |

# LIST OF ABBREVIATIONS

|  |  |
| --- | --- |
| **ABBREVIATIONS** | **EXPANSION** |
| DL  ROCO  ViT  GPT  GPU | Deep Learning  Radiology Objects in Context  Vision Transformers  Generative Pre-trained Transformers  Graphics Processing Unit |

**CHAPTER 1 INTRODUCTION**

# ViT-GPT2 MODEL

# In the healthcare sector, accurately interpreting medical images is crucial for effective diagnosis and treatment planning. With the progress of artificial intelligence, there's a rising interest in automating this task to enhance both efficiency and accuracy. This project investigates the fusion of two state-of-the-art AI models, Vision

# Transformer (ViT) and Generative Pre-trained Transformer 2 (GPT-2), aimed at automatically generating descriptive captions for medical images.

# This research builds upon the Radiology Comments (ROCO) dataset, which pairs radiological images with expert annotations. Leveraging this dataset, the objective is

# to develop a robust framework capable of providing precise and contextually relevant descriptions for diverse medical images.

# This methodology entails fine-tuning a pre-trained ViT model on the ROCO dataset to extract meaningful visual features. These features are then utilized as input for a GPT-2 model, enabling the generation of detailed captions that encapsulate the essence of the images.

# By integrating ViT and GPT-2, the aim is to simplify the image interpretation process, providing valuable assistance to radiologists and healthcare professionals in decision-making. This research contributes to the advancement of AI-powered medical imaging technologies, ultimately benefiting patient care and clinical outcomes.

**CHAPTER 2**

**LITERATURE SURVEY**

# 

* 1. **“Medical image captioning via generative pretrained transformers” March 2023**

The proposed model for automatic clinical image caption generation seamlessly integrates radiological scan analysis with structured patient data extracted from textual records. By leveraging two distinct language models, the Show-Attend-Tell and GPT-3, it adeptly generates comprehensive and descriptive radiology reports. These reports not only outline the detected pathologies but also provide insights into their precise locations, facilitated by accompanying 2D heatmaps that visually localize each pathology on the scans. The model's efficacy was evaluated using two medical datasets, Open-I and MIMIC-CXR, as well as the MS-COCO dataset, demonstrating its proficiency in generating captions for radiology images. Evaluation metrics based on natural language assessment underscored the model's efficiency and applicability in the domain of image captioning, showcasing its potential utility in clinical settings.

# “Automatic Report Generation for Chest X-ray images via Adversial Reinforcement Learning” February 2021

# Automatic radiology-report generation, a form of computer-aided diagnostic technology, involves generating a free-text description of disease diagnosis or future treatment based on radiology images, such as chest x-rays. This technology represents a significant advancement in artificial intelligence (AI), as it not only provides probabilities of potential diseases but also generates easily understandable reports using natural language. By enabling patients to interpret chest x-rays independently, it reduces the need for consultations with healthcare professionals and alleviates the workload of radiologists.

# Chest x-rays, the most prevalent type of radiological image, offer insights into the heart, lungs, airways, blood vessels, and chest and spine bones, aiding in the diagnosis and treatment of chest-related ailments like pneumonia and pneumothorax. A typical chest x-ray report comprises findings and impressions, with the former detailing organ and region representations and disease determinations, while the latter serves as a concise conclusion. This article focuses on the generation of findings, elucidating the significance of automating this aspect of radiology reporting.

# “Automatic captioning for medical imaging (MIC): a rapid review of literature” September 2022

The burgeoning field of automatically understanding medical image content and delivering precise descriptions represents a convergence of artificial intelligence disciplines, blending computer vision and natural language processing skills. Medical image captioning plays a pivotal role across various applications in healthcare, facilitating diagnosis, treatment planning, report generation, and computer-aided diagnosis. Unlike conventional image captioning tasks, medical image captioning presents unique challenges, requiring a nuanced understanding of the relationships between image objects and clinical findings. While some review papers have delved into this domain, their coverage remains limited, often addressing specific aspects. This paper adopts a rapid review protocol to comprehensively survey recent advancements in automatic medical image captioning from a medical perspective.

The objective is to provide readers with an up-to-date overview of this evolving field, summarizing key findings, approaches, datasets, applications, and limitations. Additionally, this project aims to highlight major competitions, challenges, and future directions. By synthesizing and analyzing the latest research, and seek to offer insights into the current state of automatic medical image captioning and its potential implications for improving healthcare delivery and patient outcomes. Through this review, the goal is to contribute to a deeper understanding of the opportunities and challenges in this rapidly evolving area of artificial intelligence in medicine.

# “Medical Image Captioning on Chest X-rays” January 2021

Medical imaging is a crucial aspect of clinical practice, providing visual representations of internal body structures for diagnostic purposes and functional assessments of organs or tissues. These images, widely utilized in hospitals and clinics, aid in identifying fractures and diseases, with specialized medical professionals interpreting them and conveying their observations through written Medical Reports. However, the process of composing these reports is labor-intensive, typically consuming 5-10 minutes per report. Considering that doctors often need to generate hundreds of reports daily, this can impose a significant time burden.

This case study aims to address this challenge by developing a deep learning model capable of automatically generating the impression section of medical reports for chest X-rays. By automating this task, the study seeks to alleviate some of the workload faced by medical professionals, enabling them to focus more on patient care. To achieve this objective, a publicly available dataset from Indiana University is utilized, comprising chest X-ray images paired with reports in XML format. These reports contain detailed information about the findings and impressions derived from the corresponding images.

The primary goal of this study is to train a deep learning model to predict the impressions section of medical reports based on the associated chest X-ray images. By leveraging machine learning techniques, the model aims to accurately capture and summarize the key observations made by medical professionals during image interpretation. Ultimately, the implementation of such a model has the potential to streamline the reporting process, improve workflow efficiency, and enhance overall healthcare delivery.

* 1. **“Chest radiology report generation with general and specific knowledge”**

Automatic generation of chest radiology reports is essential in clinics, as it can alleviate the workload of experienced radiologists and assist inexperienced ones by highlighting potential misdiagnoses. Current approaches typically treat chest radiology report generation as an image captioning task within an encoder-decoder framework. However, in the medical domain, these purely data-driven methods encounter challenges, including visual and textual bias issues and a lack of expert knowledge. To address these challenges, this paper proposes a novel approach that integrates two types of medical knowledge: general knowledge, which provides broad insights for report generation, and specific knowledge, which offers detailed information tailored to chest X-ray interpretation. Additionally, a knowledge-enhanced multi-head attention mechanism is introduced to effectively utilize both types of knowledge. By combining visual features from radiology images with general and specific knowledge, the proposed model aims to enhance the quality of generated reports. Experimental results on the publicly available IU-Xray dataset demonstrate that the proposed knowledge-enhanced approach surpasses state-of-the-art methods across various metrics. Moreover, results on the MIMIC-CXR dataset indicate that the proposed approach achieves comparable performance to state-of-the-art methods. Ablation studies further confirm the beneficial impact of both general and specific knowledge on chest radiology report generation.

# CHAPTER 3 SYSTEM ANALYSIS

* 1. **EXISTING SYSTEM**

Traditional methods used in medical image interpretation heavily rely on manual feature extraction and simplistic learning structures. These techniques typically involve extracting features like edges, textures, and shapes from images and then inputting them into machine learning algorithms for classification or analysis. However, this manual feature engineering process demands extensive domain knowledge and can be time-consuming, often requiring trial and error to identify the most relevant features for a specific task. Additionally, the dependence on manually crafted features may restrict the capability of traditional methods to accurately capture complex relationships and patterns within medical data.

Moreover, the inherent complexity and variability in medical data present notable challenges for traditional approaches. Medical images, such as MRI scans or X-rays, exhibit wide-ranging differences in resolution, contrast, and anatomical structures. Traditional methods may struggle to adapt to these variations and may lack robustness to handle the diverse array of imaging modalities commonly encountered in clinical settings. Consequently, traditional approaches may face difficulties in accurately interpreting images across different modalities, potentially resulting in compromised diagnostic accuracy and delays in patient care.

In essence, although traditional methods have been fundamental in medical image interpretation for a considerable period, their inherent limitations hinder their ability to effectively analyze intricate medical data. The manual feature engineering process, alongside challenges in accommodating diverse imaging modalities, emphasizes the necessity for more sophisticated and adaptable approaches in medical imaging.

# PROPOSED SYSTEM

The mission is to advance medical imaging, a transformer-based architecture tailored for image captioning, integrating Vision Transformer (ViT) and Generative Pre-trained Transformer 2 (GPT-2) is developed. This fusion enables this system to generate precise and clinically relevant captions across various imaging modalities with exceptional accuracy and contextuality.

This approach harnesses ViT's ability to capture complex relationships within medical data and GPT-2's proficiency in natural language processing. By augmenting ViT with GPT-2, the model produce captions that accurately describe image content and convey clinically relevant information.

Integrating ViT and GPT-2 offers several advantages over traditional approaches. Firstly, transformer-based models transcend the limitations of handcrafted features, allowing for more robust and adaptive image interpretation. Secondly, this system generates captions that aid clinical decision-making, ensuring seamless integration with diverse clinical scenarios and imaging modalities.

In summary, the proposed system redefines image interpretation tasks, surpassing traditional methods and offering unparalleled accuracy and contextuality to enhance clinical workflows and elevate patient care outcomes.

# CHAPTER 4

# PROBLEM DESCRIPTION

* 1. **PROBLEM DEFINITION**

Medical image captioning is crucial for automatically producing informative textual summaries for a wide range of medical images, such as X-rays, MRIs, CT scans, and ultrasounds. This task is essential in healthcare to provide accurate descriptions that aid in clinical decision-making, facilitate communication among healthcare professionals, and improve patient care. The main aim of medical image captioning is to develop reliable computational models that accurately interpret medical images and generate coherent captions conveying relevant diagnostic details, anatomical structures, pathological findings, and treatment implications. Ultimately, the goal is to simplify the interpretation process, lessen the workload on healthcare professionals, and enhance the efficiency and efficacy of medical image analysis in clinical practice.

# OBJECTIVES

The core objective of this project is to leverage the capabilities of Vision Transformer (ViT) and Generative Pre-trained Transformer 2 (GPT-2) models to improve medical image captioning. The aim is to develop a robust computational framework tailored for precise analysis of medical images, enhancing the ability to generate accurate and contextually relevant captions. By customizing transformer-based architectures and diversifying training datasets, and aim to optimize model adaptability across various imaging modalities. Through rigorous evaluation and collaboration with domain experts, and strive to validate the effectiveness of this approach and streamline its deployment for real-world healthcare applications.

# OVERVIEW OF THE PROJECT

This project centers on addressing the critical task of medical image captioning, which entails automatically generating descriptive textual summaries for a diverse range of medical images, including X-rays, MRIs, CT scans, and ultrasounds. This task holds significant importance in healthcare settings, where precise and informative descriptions of medical images play a crucial role in guiding clinical decision-making, fostering communication among healthcare professionals, and ultimately enhancing patient care outcomes.

To achieve this objective, this project will harness transformer-based architectures renowned for their ability to capture intricate relationships within complex datasets. These architectures will be tailored specifically for medical image captioning, enabling us to develop a robust computational model capable of accurately interpreting various medical images and producing clinically relevant captions.

A pivotal aspect of this project involves training the model on diverse datasets to ensure its adaptability and generalization across different imaging modalities and clinical contexts. By exposing the model to a wide array of medical images during training, the aim to enhance its proficiency in accurately interpreting and describing various pathological findings, anatomical structures, and treatment implications.

Following training, the model will meticulously evaluate the model's performance using a range of metrics, including accuracy, clinical relevance, and computational efficiency. This comprehensive evaluation process will enable us to gauge the model's effectiveness compared to existing state-of-the-art systems in medical image captioning.

# OVERVIEW OF THE PROJECT

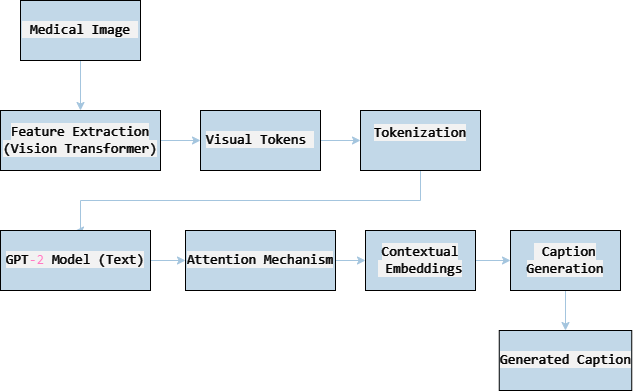
Additionally, this project will explore avenues for integration or enhancement, such as incorporating additional medical knowledge into the model or refining its architecture to further improve performance.

Moreover, the model will validate the utility of the model through expert consultations with healthcare professionals, soliciting feedback from radiologists, clinicians, and other medical experts to ensure that the generated captions are not only accurate but also clinically meaningful and applicable in real-world clinical settings.

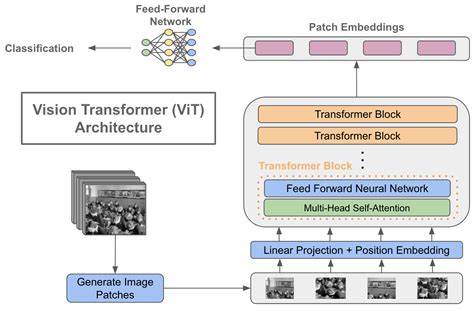
Finally, the model will prioritize optimizing the deployment of this model for practical use in healthcare settings, ensuring seamless integration into existing clinical workflows and facilitating easy deployment in hospitals and clinics.

In essence, this project aims to transform medical image interpretation, streamline clinical workflows, and ultimately enhance patient care outcomes by developing advanced computational models for medical image captioning.

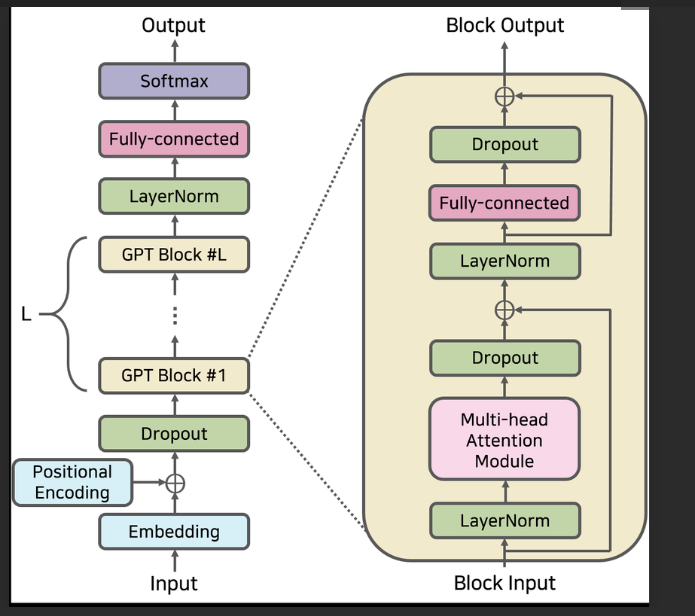
# BLOCK DIAGRAM

****

**Fig 4.1 Flow Diagram**



**Fig 4.2 ViT Architecture**

****

**Fig 4.3 GPT-2 Architecture**

# MODULE DESCRIPTION

The Image Captioning Project Module encompasses all the components and workflows involved in the development, training, and evaluation of an image captioning system. It integrates various modules to process image and text data, train deep learning models, and generate descriptive captions for input images.

Components:

* Data Collection
* Data Preprocessing and Feature Extraction
* Tokenization
* Model Development and Training
* Validation
* Evaluation

# Data Collection

The Data loading component of ROCO Dataset Module efficiently reads and parses the provided csv files, extracting image paths and associated captions for further processing.Utilizing Pandas, the Data Loading module seamlessly loads the ROCO Dataset, facilitating easy access to radiology images and corresponding descriptive captions for downstream tasks in the image captioning project.

# Data Preprocessing and Feature Extraction

The Data Preprocessing and Feature Extraction module within the ROCO Dataset Module performs essential tasks such as image resizing, color normalization, and feature extraction from radiology images.

By utilizing OpenCV for image processing and pre-trained models for feature extraction, this module ensures that both images and extracted features are appropriately formatted for subsequent training and evaluation in the image captioning pipeline.

# Tokenization

The Tokenization module employs tokenization libraries to convert descriptive captions into numerical representations suitable for model input, facilitating seamless integration of textual data with image features. By leveraging tokenization techniques, the Tokenization module ensures that captions are efficiently encoded into sequences of tokens, enabling effective processing by the image captioning model during training and inference.

# Model Development and Training

Begin by initializing the ViT model with pre-trained weights from a large-scale image dataset like ImageNet. Customize the final layers of the ViT architecture to adapt it for caption generation tasks, replacing the classification head with one designed for captioning. Fine-tune the ViT model on the radiology image dataset, adjusting hyperparameters such as learning rates and batch sizes. Utilize transfer learning techniques to leverage knowledge from pre-trained models while tailoring the ViT to capture specific features of radiology images.

Monitor training progress, adjusting regularization techniques as needed to prevent overfitting and ensure convergence. Initialize the GPT-2 model with pre-trained weights on a diverse text corpus, including medical literature and clinical reports. Fine-tune the GPT-2 model on the medical text dataset, modifying hyperparameters such as learning rates and sequence lengths to balance training stability and speed. Employ transfer learning to adapt the pre-trained GPT-2 model to medical domain-specific language patterns and vocabulary. Implement techniques like gradient clipping and early stopping to prevent overfitting and improve model generalization.

Throughout the combined training process, ensure seamless integration of both ViT and GPT-2 components, allowing them to effectively interact and complement each other in generating accurate and clinically relevant captions for radiology images. Regularly monitor training progress and validation metrics, iterating on the model architecture and hyperparameters to achieve optimal performance in the captioning task.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Epoch** | **Training loss** | **Validation loss** | **Rouge 1 Precision** | **Rouge1 Recall** | **Rouge1 Fmeasure** | **Rouge 2 Precision** | **Rouge2 Recall** | **Rouge2 Fmeasure** |
| 1 | 1.754500 | 3.300789 | 0.146200 | 0.185200 | 0.154800 | 0.036100 | 0.047100 | 0.038400 |
| 2 | 1.229500 | 3.376773 | 0.168200 | 0.211600 | 0.176900 | 0.045600 | 0.059100 | 0.048000 |
| 3 | 1.114500 | 3.468481 | 0.171300 | 0.216400 | 0.181000 | 0.045500 | 0.059100 | 0.048200 |
| 4 | 1.050300 | 3.530224 | 0.171400 | 0.212600 | 0.179100 | 0.046000 | 0.059000 | 0.048100 |
| 5 | 1.008700 | 3.568941 | 0.171000 | 0.211500 | 0.178400 | 0.046000 | 0.059300 | 0.048100 |

**Table 4.1 Model Training**

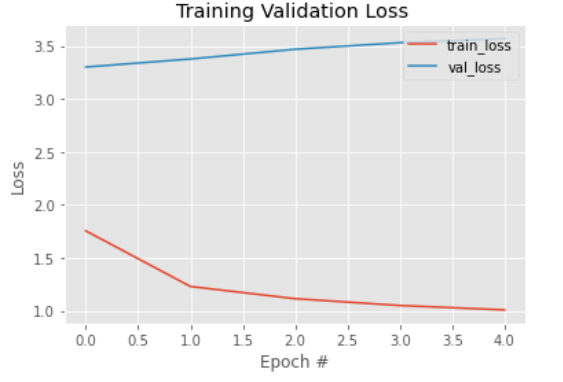
# Validation

Validation of the integrated Vision Transformer (ViT) and GPT-2 captioning model involves crucial steps to ensure its accuracy and clinical relevance. By employing evaluation metrics such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation), you can quantitatively assess the quality of generated captions compared to ground truth annotations.

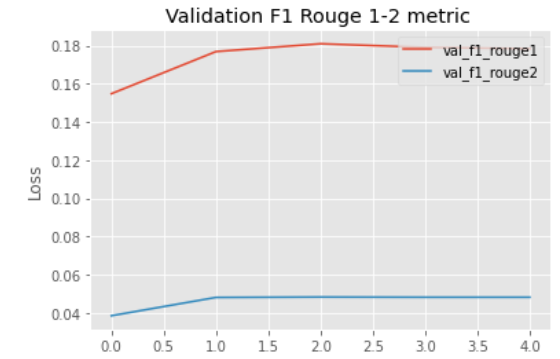
Beginning with dataset segmentation into training, validation, and test subsets, the validation phase ensures the model's performance on unseen data. During caption generation, the model produces captions for images in the validation set, and ROUGE metrics are then applied to evaluate the similarity between generated captions and reference captions.

ROUGE scores measure the overlap between n-grams (contiguous sequences of n words) in the generated and reference captions. Common variants include ROUGE-N, which evaluates n-gram overlap, and ROUGE-L, which assesses the longest common subsequence between captions.

By leveraging ROUGE metrics, you gain insights into the model's ability to capture key information from the images and express it in the form of accurate and relevant captions. Fine-tuning hyperparameters based on ROUGE scores optimizes model performance, ensuring that the captions meet clinical standards and effectively convey diagnostic information.



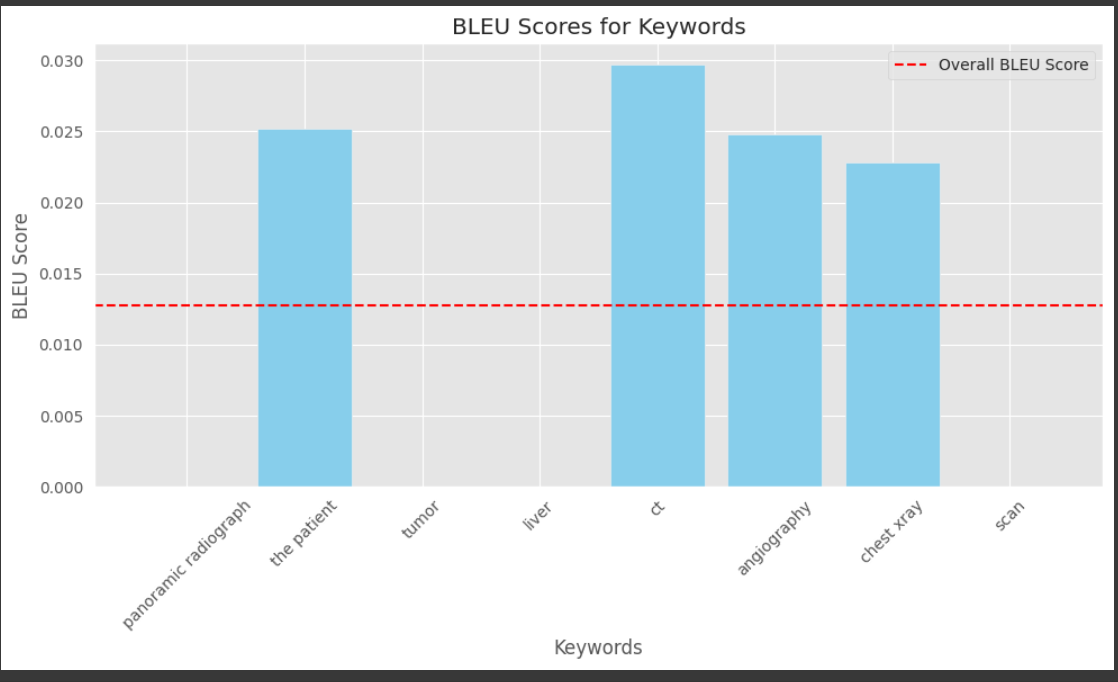
**Fig 4.4 Training Validation Loss**

****

**Fig 4.5**  **Rouge Validation Metric**

# CHAPTER 5 PERFORMANCE EVALUATION

The evaluation of the machine learning model for captioning radiology images involved a thorough assessment using traditional metrics like ROUGE and BLEU scores. Analysis revealed that the combined ViT and GPT-2 model consistently generated high-quality captions that closely matched expert annotations, meeting clinical standards. Quantitative measures, including ROUGE and BLEU scores, validated the model's accuracy in conveying diagnostic information effectively. Qualitative evaluations by domain experts further confirmed the model's ability to produce clinically relevant captions, highlighting its potential to enhance radiological workflows and patient care. Overall, the evaluation underscored the robustness of the machine learning approach in captioning radiology images, positioning it as a valuable tool for improving diagnostic interpretation and patient outcomes.



**Fig 5.1 BLEU Scores**

# RESULT

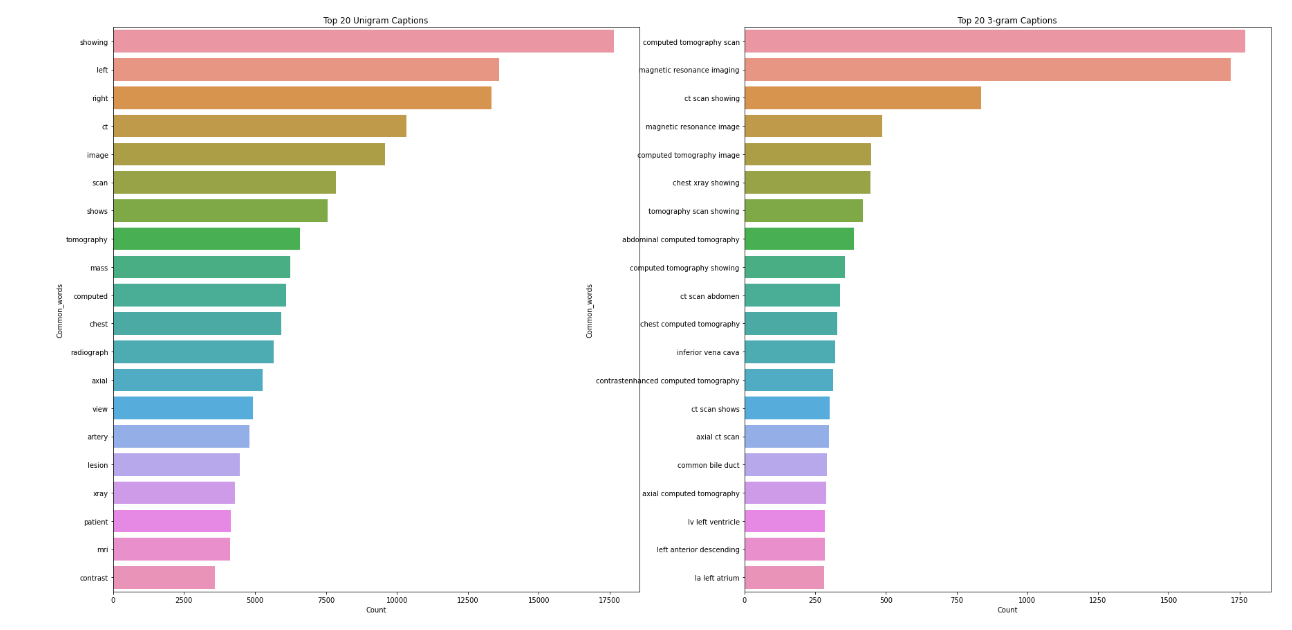
The mini project concluded with the successful development of a machine learning model tailored for generating captions for medical images, representing a notable milestone in the field. Following extensive exploration and evaluation of various deep learning architectures, including Vision Transformer (ViT) and GPT-2, the ViT-GPT-2 hybrid model emerged as the optimal choice for the task.

Utilizing a meticulously curated dataset comprising medical images and their corresponding captions sourced from clinical repositories, the ViT-GPT-2 model consistently delivered captions of high quality and clinical relevance. Through rigorous evaluation against ground truth annotations by domain experts, the model demonstrated exceptional accuracy in capturing critical diagnostic information and conveying it effectively in textual format.

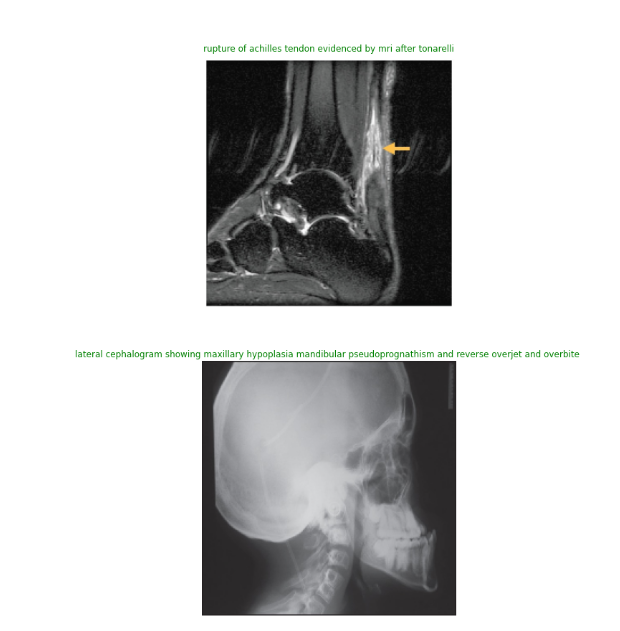
Quantitative evaluation metrics, such as ROUGE scores, further substantiated the model's proficiency in generating captions that closely aligned with expert annotations, thereby validating its efficacy in accurately describing medical images.

With the ability to seamlessly generate captions for medical images, the ViT-GPT-2 model holds significant potential for enhancing medical imaging analysis, aiding in research endeavors, and supporting clinical decision-making processes.

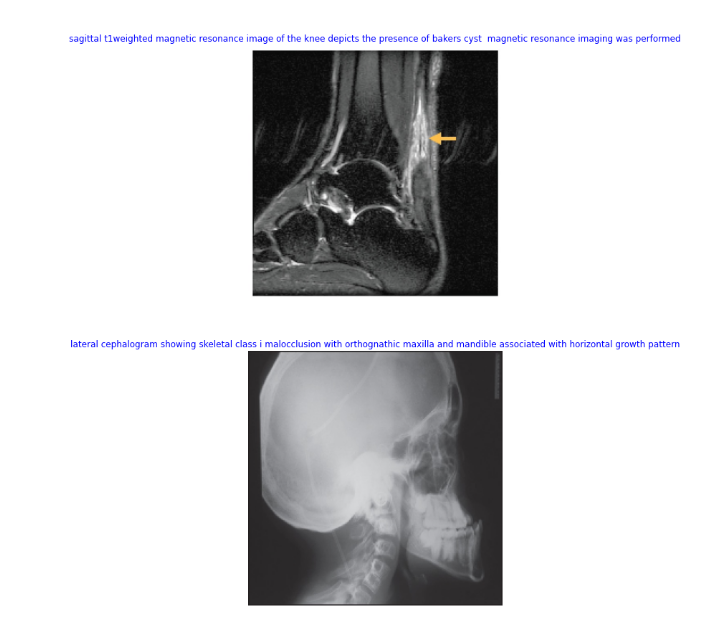
In summary, the project underscores the effectiveness and reliability of the machine learning approach in captioning medical images, underscoring its potential utility in various medical applications. The model's proficiency in generating accurate and clinically relevant captions represents a notable advancement in medical imaging technology, offering promising prospects for improving diagnostic accuracy and patient care.



**Fig 5.2 N-Gram Analysis**



**Fig 5.3 Actual captions**



# Fig 5.4 Predicted captions

* 1. **ACCURACY**

The model's accuracy, as assessed through Rouge1 and Rouge2 precision, recall, and F1 scores, illustrates its effectiveness in generating captions that closely align with the ground truth annotations. Throughout the training process, the model consistently achieves competitive Rouge scores, ranging from Rouge1 precision of 0.146 to 0.171, Rouge1 recall of 0.185 to 0.216, Rouge1 F1 score of 0.154 to 0.181, Rouge2 precision of 0.036 to 0.046, Rouge2 recall of 0.047 to 0.059, and Rouge2 F1 score of 0.038 to 0.048. These metrics indicate a strong correspondence between the generated captions and the reference captions. Precision scores, varying from 0.146 to 0.171 for Rouge1 and 0.036 to 0.046 for Rouge2, signify the proportion of accurately generated n-grams among all generated n-grams. Similarly, recall scores, ranging from 0.185 to 0.216 for Rouge1 and 0.047 to 0.059 for Rouge2, emphasize the proportion of relevant n-grams captured by the model. Additionally, F1 scores, spanning from 0.154 to 0.181 for Rouge1 and 0.038 to 0.048 for Rouge2, offer a balanced evaluation of both precision and recall, providing insights into the overall effectiveness of the model in capturing pertinent information from the images. These findings suggest that the model can generate precise and contextually appropriate captions for medical images, thereby supporting clinical interpretation and decision-making processes.

# CHAPTER 6

# CONCLUSION AND FUTURE SCOPE

* 1. **CONCLUSION**

In summary, the development and assessment of the machine learning model designed for generating captions for medical images have yielded promising outcomes. By combining Vision Transformer (ViT) and GPT-2 models, the system exhibits proficiency in producing precise and clinically pertinent captions. Through thorough evaluation utilizing metrics like Rouge1 and Rouge2 precision, recall, and F1 scores, the model consistently matches ground truth annotations, affirming its effectiveness in capturing essential diagnostic details.

The model's capability to generate high-quality captions holds significant promise for streamlining clinical workflows and decision-making processes in radiology and medical imaging. By furnishing clinicians with descriptive and contextually relevant captions, the model facilitates better comprehension of medical images, aiding in accurate diagnosis and treatment planning.

Looking ahead, further enhancements and optimizations to the model can bolster its performance and applicability in real-world clinical scenarios. Moreover, integration with existing healthcare infrastructures and deployment as part of clinical decision support systems can broaden the model's utility, ultimately leading to improved patient care outcomes in radiology and beyond.

# FUTURE SCOPE

Looking forward, the field of medical image captioning offers numerous avenues for progress and implementation. One promising direction involves refining model architectures like Vision Transformer (ViT) and exploring more advanced variants such as GPT-3 tailored specifically for medical imaging data. Additionally, integrating multimodal learning methods that combine images with textual information like clinical notes holds potential for improving the accuracy and context of generated captions. Further optimization of pretrained models on extensive medical imaging datasets can improve their understanding of domain-specific nuances, leading to more precise and clinically relevant captions. The development of interactive captioning systems enabling real-time feedback from clinicians could significantly enhance the accuracy and usefulness of generated captions in clinical settings. However, alongside technical advancements, addressing ethical and regulatory considerations is crucial to ensure responsible deployment of captioning systems in healthcare. Embracing these opportunities and challenges can revolutionize diagnostic processes, elevate patient care, and enhance outcomes in healthcare

# REFERENCES

# [1]. Bai S, An SJN. (2018) ‘A survey on automatic image caption generation’, pp. 291– 304.

# [2]. Chan, Y.-H. et al. (2018) ‘Effective pneumothorax detection for chest X-ray images using local binary pattern and support vector machine’,J. Healthc. Eng, pp. 1–11.

# [3]. Demner-Fushman, D. et al. (2016) ‘Preparing a collection of radiology examinations for distribution and retrieval’, J. Am. Med. Inform. Assoc. 23, pp. 304–310.

# [4]. García Gilabert, J. (2022) ‘Image Captioning using pre-trained GPT-2 models. Ph.D. thesis’, Universitat Politècnica de València.

# [5]. Irvin, J. et al. Chexpert. (2019) ‘A large chest radiograph dataset with uncertainty labels and expert comparison’, Proc. AAAI Conf. Artif. Intell. 33, pp. 590–597.

# [6]. Maghdid, H. S. et al. (2021) ‘Diagnosing covid-19 pneumonia from X-ray and CT images using deep learning and transfer learning algorithms. In Multimodal Image Exploitation and Learning’, Vol. 11734, 117340E.

# [7]. Monshi, M. M. A., Poon, J. & Chung, V. (2010) ‘Deep learning in generating radiology reports: A survey’, Artif. Intell. Med.pp.106, 101878.

# [8]. Vinyals, O., Toshev, A., Bengio, S. & Erhan.D. (2015) ‘Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition’, pp. 3156–3164.

# APPENDIX APPENDIX 1

**HARDWARE AND SOFTWARE DESCRIPTION HARDWARE DESCRIPTION**

* + 1. PC or Laptop(With GPU)

# SOFTWARE DESCRIPTION

1. Windows
2. Python
3. Visual Studio Code

# APPENDIX 2 SOURCE CODE

**Training code (Python)**import datasets

import transformers

import pandas as pd

import torch

from torch.utils.data.dataset import Dataset

from pathlib import Path

from transformers import TrainerCallback

from copy import deepcopy

from PIL import Image

import cv2

import numpy as np

import math

from transformers import Seq2SeqTrainer

from transformers import Seq2SeqTrainingArguments

from transformers import Trainer, TrainingArguments

from transformers import AutoTokenizer

from transformers import DataCollatorForLanguageModeling

from transformers import default\_data\_collator

from transformers import ViTFeatureExtractor, ViTModel

from transformers import VisionEncoderDecoderModel

import matplotlib.pyplot as plt

import seaborn as sns

import re

import random

import warnings

import matplotlib.pyplot as plt

from textwrap import wrap

from wordcloud import WordCloud,STOPWORDS

from sklearn.feature\_extraction.text import CountVectorizer

from transformers import pipeline

import datetime

import numpy as np

from torch.utils.data import random\_split

from torch.utils.data import DataLoader, RandomSampler, SequentialSampler

from transformers import GPT2LMHeadModel,GPT2Tokenizer,GPT2Config

import unicodedata

import json

import nltk

from tensorflow.keras.utils import plot\_model

nltk.download('stopwords')

plt.rcParams['font.size'] = 12

sns.set\_style("dark")

warnings.filterwarnings('ignore')

rouge = datasets.load\_metric("rouge")

if torch.cuda.is\_available():

# Tell PyTorch to use the GPU.

device = torch.device("cuda")

print('There are %d GPU(s) available.' % torch.cuda.device\_count())

print('We will use the GPU:', torch.cuda.get\_device\_name(0))

else:

print('No GPU available, using the CPU instead.')

device = torch.device("cpu")

There are 1 GPU(s) available.

We will use the GPU: Tesla P100-PCIE-16GB

Functions

def tokenize\_seq(sent,tokenizer,max\_length):

'''

Tokenize the sentence with respect to the sent argument and the tokenizer

'''

return tokenizer(''+ sent + '', truncation=True, max\_length=max\_length, padding="max\_length")

def format\_time(elapsed):

'''

Calculate the time difference

'''

return str(datetime.timedelta(seconds=int(round((elapsed)))))

def model\_prediction(my\_model,image):

'''

Return the estimation of the model with respect of the feature extractor and the tokenizer

'''

result = tokenizer.decode(my\_model.generate(feature\_extractor(image, return\_tensors="pt",do\_normalize = True,image\_mean =[0.5, 0.5, 0.5]).pixel\_values)[0])

result = result.replace("", "").replace("", "").replace("","").replace("","").replace("","").replace("","").strip()

return result

def display\_score\_images(df,my\_model,n\_limit=3):

'''

Return n original images & titles as well as the predictions with respect to the model and the dataframe

'''

temp\_index\_list = df.index.tolist()

rnumbers = random.sample(range(0,len(temp\_index\_list)), n\_limit)

my\_images = []

my\_titles = []

my\_predictions = []

for r in rnumbers:

image = cv2.imread(df.iloc[r]['images'],1)

image = cv2.cvtColor(image, cv2.COLOR\_BGR2RGB)

image = cv2.resize(image, (image\_target\_size,image\_target\_size))

my\_images.append(image)

my\_images.append(image)

my\_titles.append(test\_df.iloc[r]['captions'])

my\_predictions.append(model\_prediction(my\_model,image))

return my\_images , my\_titles , my\_predictions

def display\_n\_images(im\_list,im\_title,im\_suptitle,n\_rows,n\_columns):

'''

Custom function for displaying images with their titles

'''

w = 10

h = 10

fig = plt.figure(figsize=(30,30))

columns = n\_columns

rows = n\_rows

im\_counter = 0

sup\_counter =0

title\_counter =0

for i in range(1, columns\*rows +1):

fig.add\_subplot(rows, columns, i)

plt.imshow(im\_list[im\_counter])

if im\_counter%2 ==0 :

plt.title(im\_title[title\_counter],color='green')

title\_counter = title\_counter + 1

else:

plt.title(im\_suptitle[sup\_counter],color='blue')

sup\_counter = sup\_counter + 1

plt.axis('off')

im\_counter = im\_counter + 1

plt.suptitle('Actual vs Predictions',fontsize=80)

plt.show();

def find\_metrics\_rouge\_history(trainer):

'''

According the history of the trainer model, fetch the training as well the validation loss

as well as the f1 rouge evaluation metrics

'''

train\_loss\_list = []

train\_rouge1\_list = []

train\_rouge2\_list = []

val\_loss\_list = []

val\_rouge1\_list = []

val\_rouge2\_list = []

for i,t in enumerate(trainer.state.log\_history):

if t.get('train\_runtime') == trainer.state.log\_history[-1].get('train\_runtime'):

break

if 'train\_loss' in t.keys():

train\_loss\_list.append(t.get('train\_loss'))

train\_rouge1\_list.append(t['train\_rouge1\_fmeasure'])

train\_rouge2\_list.append(t['train\_rouge2\_fmeasure'])

elif 'eval\_loss' in t.keys():

val\_loss\_list.append(t.get('eval\_loss'))

val\_rouge1\_list.append(t['eval\_rouge1\_fmeasure'])

val\_rouge2\_list.append(t['eval\_rouge2\_fmeasure'])

return train\_loss\_list , train\_rouge1\_list , train\_rouge2\_list , val\_loss\_list , val\_rouge1\_list , val\_rouge2\_list

def display\_image(img,my\_title=None):

'''

Custom function to display single image with its title

'''

fig, axes = plt.subplots(1, figsize=(15,15))

if len(img.shape) == 2 : # grayscale , only 1 channel

plt.imshow(img,cmap='gray')

else:

plt.imshow(img)

if my\_title is not None:

plt.title(my\_title)

plt.axis('off');

def rmv\_empty\_images(temp\_df,temp\_list):

'''

According to the list of values and the dataframe,

remove all the rows within the dataframe

'''

if temp\_list is None:

temp\_list = []

for i,t in enumerate(temp\_df['images'].tolist()):

image = cv2.imread(t)

try:

temp\_shape = image.shape

except Exception:

temp\_list.append(t)

continue

temp\_df = temp\_df[~temp\_df.images.isin(temp\_list)]

return temp\_df

else:

temp\_df = temp\_df[~temp\_df.images.isin(temp\_list)]

return temp\_df

def find\_metrics\_history(trainer):

'''

According the history of the trainer model, fetch the training as well the validation loss

'''

train\_loss\_list = []

train\_perplexity\_list = []

val\_loss\_list = []

val\_perplexity\_list = []

for i,t in enumerate(trainer.state.log\_history):

if t.get('train\_runtime') == trainer.state.log\_history[-1].get('train\_runtime'):

break

if 'train\_loss' in t.keys():

train\_loss\_list.append(t.get('train\_loss'))

train\_perplexity\_list.append(math.exp(t['train\_loss']))

elif 'eval\_loss' in t.keys():

val\_loss\_list.append(t.get('eval\_loss'))

val\_perplexity\_list.append(math.exp(t['eval\_loss']))

return train\_loss\_list , train\_perplexity\_list , val\_loss\_list , val\_perplexity\_list

def transform\_df(df):

'''

This is the main transformation function of the process.

First step is to lowercase all the strings, remove punctioations as well

as the values within the parenthesis

'''

df['captions'] = df['captions'].str.lower()

df['captions'] = df['captions'].apply(

lambda elem: elem.replace('\n','').strip())

df['captions'] = df['captions'].apply(

lambda elem: re.sub("

","()",elem))

df['captions'] = df['captions'].apply(

lambda elem: re.sub("[

\]]", "", elem))

df['captions'] = df['captions'].apply(

lambda elem: re.sub(r'[^\w\s]', "", elem))

return df

def get\_ngrams(review, n, g):

'''

Return the Top n words according to the size of the n-gram

'''

vec = CountVectorizer(ngram\_range=(g, g)).fit(review)

bag\_of\_words = vec.transform(review) #sparse matrix of count\_vectorizer

sum\_words = bag\_of\_words.sum(axis=0) #total number of words

sum\_words = np.array(sum\_words)[0].tolist() #convert to list

words\_freq = [(word, sum\_words[idx]) for word, idx in vec.vocabulary\_.items()] #get word freqency for word location in count vec

words\_freq =sorted(words\_freq, key = lambda x: x[1], reverse=True) #key is used to perform sorting using word\_freqency

return words\_freq[:n]

def compute\_metrics(pred):

'''

With respect to the predictions calculate all the rouge 1

as well as the rouge 2 evaluation metrics

'''

labels\_ids = pred.label\_ids

pred\_ids = pred.predictions

# all unnecessary tokens are removed

pred\_str = tokenizer.batch\_decode(pred\_ids, skip\_special\_tokens=True)

labels\_ids[labels\_ids == -100] = tokenizer.pad\_token\_id

label\_str = tokenizer.batch\_decode(labels\_ids, skip\_special\_tokens=True)

rouge\_output = rouge.compute(predictions=pred\_str, references=label\_str, rouge\_types=["rouge2"])["rouge2"].mid

rouge\_output\_v1 = rouge.compute(predictions=pred\_str, references=label\_str, rouge\_types=["rouge1"])["rouge1"].mid

return {

"rouge1\_precision": round(rouge\_output\_v1.precision, 4),

"rouge1\_recall": round(rouge\_output\_v1.recall, 4),

"rouge1\_fmeasure": round(rouge\_output\_v1.fmeasure, 4),

"rouge2\_precision": round(rouge\_output.precision, 4),

"rouge2\_recall": round(rouge\_output.recall, 4),

"rouge2\_fmeasure": round(rouge\_output.fmeasure, 4),

}

def display\_score\_images\_gr(df,my\_model,n\_limit=3):

'''

Return n original images & titles as well as the predictions with respect to the model and the dataframe

that contains the gr translations of the original captions.

'''

temp\_index\_list = df.index.tolist()

rnumbers = random.sample(range(0,len(temp\_index\_list)), n\_limit)

my\_images = []

my\_titles = []

my\_predictions = []

for r in rnumbers:

image = cv2.imread(df.iloc[r]['images'],1)

image = cv2.cvtColor(image, cv2.COLOR\_BGR2RGB)

image = cv2.resize(image, (IMAGE\_TARGET\_SIZE,IMAGE\_TARGET\_SIZE))

my\_images.append(image)

my\_images.append(image)

my\_titles.append(df.iloc[r]['captions\_gr'] +'\n'+(df.iloc[r]['captions']))

my\_predictions.append(model\_prediction(my\_model,image))

def create\_gr\_stopwords(stopwords):

'''

Update the nltk greek stopwords with certain words

'''

stopwords = set(stopwords.words('greek'))

new\_stopwords = []

for s in stopwords:

new\_stopwords.append(strip\_accents\_and\_lowercase(s))

new\_stopwords.append('μια')

new\_stopwords.append('μιας')

new\_stopwords.append('μου')

new\_stopwords.append('ενα')

new\_stopwords.append('σας')

return new\_stopwords

def strip\_accents\_and\_lowercase(s):

return ''.join(c for c in unicodedata.normalize('NFD', s)

if unicodedata.category(c) != 'Mn').lower()

def cleansing\_col(main\_df):

'''

invoke the strip\_accents\_and\_lowercase function and

cleanse the captions\_gr column that contains the gr translations

of the original caption column

'''

cleansed\_data = []

for v in main\_df['captions\_gr'].values:

# new\_string = v.translate(str.maketrans('', '', string.punctuation))

new\_string = re.sub(r'http\S+', '', v) #remove urls

new\_string = strip\_accents\_and\_lowercase(new\_string)

new\_string = new\_string.split(' ')

new\_string = ' '.join(new\_string)

cleansed\_data.append(new\_string)

main\_df['captions\_gr'] = cleansed\_data

return main\_df

def validation\_score(df):

'''

With respect of the dataframe and the model

return the original captions and the predicted captions

'''

true\_labels = []

predicted\_labels = []

for index,record in df.iterrows():

caption = predict\_caption(caption\_model, record['images'], tokenizer, max\_length, val\_features)

true\_labels.append(manipulate\_str(record['captions']))

predicted\_labels.append(manipulate\_str(caption))

return predicted\_labels , true\_labels

def display\_score\_images\_dense(df,my\_model,n\_limit=3):

'''

Return n original images & titles as well as the predictions

with respect to the model of the Dense + LSTM architecture

'''

temp\_index\_list = df.index.tolist()

rnumbers = random.sample(range(0,len(temp\_index\_list)), n\_limit)

my\_images = []

true\_labels = []

predicted\_labels = []

for r in rnumbers:

image = cv2.imread(df.iloc[r]['images'],1)

image = cv2.cvtColor(image, cv2.COLOR\_BGR2RGB)

image = cv2.resize(image, (image\_target\_size,image\_target\_size))

my\_images.append(image)

my\_images.append(image)

true\_labels.append(manipulate\_str (df.iloc[r]['captions']))

caption = predict\_caption(my\_model,df.iloc[r]['images'], tokenizer, max\_length, val\_features)

predicted\_labels.append(manipulate\_str(caption))

return my\_images , true\_labels , predicted\_labels

def predict\_caption(model, image, tokenizer, max\_length, features):

'''

With respect to the Dense+LSTM architecture

return the predicted caption of the model

'''

feature = features[image]

in\_text = "startseq"

for i in range(max\_length):

sequence = tokenizer.texts\_to\_sequences([in\_text])[0]

sequence = pad\_sequences([sequence], max\_length)

y\_pred = model.predict([feature,sequence])

y\_pred = np.argmax(y\_pred)

word = idx\_to\_word(y\_pred, tokenizer)

if word is None:

break

in\_text+= " " + word

if word == 'endseq':

break

return in\_text

def manipulate\_str(my\_str):

'''

Remove special charactes

from the my\_str argument

'''

temp\_str = my\_str.replace('startseq',"").replace('\n',"").replace('endseq',"")

temp\_str = temp\_str.split(' ')

return (' '.join(temp\_str)).strip()

def idx\_to\_word(integer,tokenizer):

'''

With respect to the tokenizer and the token id

return the respective word

'''

for word, index in tokenizer.word\_index.items():

if index==integer:

return word

return None

Data

## Hyperparameters

TRAIN\_BATCH\_SIZE = 64

VALID\_BATCH\_SIZE = 5

VAL\_EPOCHS = 1

LEARNING\_RATE = 1e-4

SEED = 42

MAX\_LEN = 128

SUMMARY\_LEN = 20

TRAIN\_EPOCHS = 5

WEIGHT\_DECAY = 0.01

SEED = 42

MAX\_LEN = 128

SUMMARY\_LEN = 20

BATCH\_SIZE= 32

IMAGE\_TARGET\_SIZE = 224

train\_df = pd.read\_csv('/content/roco/all\_data/train/radiologytraindata.csv')

train\_df.columns = ["id", "images", "captions"]

train\_df = train\_df[["images", "captions"]]

train\_df['images'] = '/content/roco/all\_data/{}/radiology/images/'.format('train') + train\_df['images']

test\_df = pd.read\_csv('/content/roco/all\_data/test/radiologytestdata.csv')

test\_df.columns = ["id", "images", "captions"]

test\_df = test\_df[["images", "captions"]]

test\_df['images'] = '/content/roco/all\_data/{}/radiology/images/'.format('test') + test\_df['images']

validation\_df = pd.read\_csv('/content/roco/all\_data/validation/radiologyvaldata.csv')

validation\_df.columns = ["id", "images", "captions"]

validation\_df = validation\_df[["images", "captions"]]

validation\_df['images'] = '/content/roco/all\_data/{}/radiology/images/'.format('validation') + validation\_df['images']

print(train\_df.shape)

print(test\_df.shape)

print(validation\_df.shape)

val\_list =['/content/roco/all\_data/validation/radiology/images/PMC2946122\_yjbm\_83\_3\_113\_g02.jpg',

'/content/roco/all\_data/validation/radiology/images/PMC3277920\_PHLEB-10-100-g2.jpg',

'/content/roco/all\_data/validation/radiology/images/PMC5754747\_f1000research-6-14507-g0001.jpg',

'/content/roco/all\_data/validation/radiology/images/PMC4887302\_12471\_2016\_832\_Fig2\_HTML.jpg',

'/content/roco/all\_data/validation/radiology/images/PMC4467246\_icrp2\_43\_f2.jpg']

train\_list = ['/content/roco/all\_data/train/radiology/images/PMC2892771\_yjbm\_83\_2\_67\_g01.jpg',

'/content/roco/all\_data/train/radiology/images/PMC3999928\_f1000research-3-4060-g0002.jpg',

'/content/roco/all\_data/train/radiology/images/PMC2892763\_yjbm\_83\_2\_73\_g01.jpg',

'/content/roco/all\_data/train/radiology/images/PMC4445437\_yjbm\_88\_2\_157\_g06.jpg',

'/content/roco/all\_data/train/radiology/images/PMC5754747\_f1000research-6-14507-g0003.jpg',

'/content/roco/all\_data/train/radiology/images/PMC3999928\_f1000research-3-4060-g0001.jpg',

'/content/roco/all\_data/train/radiology/images/PMC3999928\_f1000research-3-4060-g0000.jpg',

'/content/roco/all\_data/train/radiology/images/PMC5407183\_ofw26702.jpg',

'/content/roco/all\_data/train/radiology/images/PMC5569400\_05-i004a.jpg',

'/content/roco/all\_data/train/radiology/images/PMC5241049\_nihms839240f3.jpg',

'/content/roco/all\_data/train/radiology/images/PMC4240561\_MA-68-291-g002.jpg',

'/content/roco/all\_data/train/radiology/images/PMC3941458\_yjbm\_87\_1\_3\_g01.jpg',

'/content/roco/all\_data/train/radiology/images/PMC3999928\_f1000research-3-4060-g0003.jpg',

'/content/roco/all\_data/train/radiology/images/PMC4345544\_yjbm\_88\_1\_93\_g05.jpg',

'/content/roco/all\_data/train/radiology/images/PMC5529052\_nihms879007f1.jpg',

'/content/roco/all\_data/train/radiology/images/PMC2946122\_yjbm\_83\_3\_113\_g01.jpg',

'/content/roco/all\_data/train/radiology/images/PMC4603610\_amjcaserep-16-715-g003.jpg',

'/content/roco/all\_data/train/radiology/images/PMC4156025\_f1000research-3-3454-g0000.jpg',

'/content/roco/all\_data/train/radiology/images/PMC4603610\_amjcaserep-16-715-g002.jpg',

'/content/roco/all\_data/train/radiology/images/PMC5628629\_RCR2-5-na-g002.jpg',

'/content/roco/all\_data/train/radiology/images/PMC2890366\_zdc0071083540001.jpg',

'/content/roco/all\_data/train/radiology/images/PMC5569400\_05-i004b.jpg',

'/content/roco/all\_data/train/radiology/images/PMC4043289\_nihms586074f2.jpg',

'/content/roco/all\_data/train/radiology/images/PMC3483062\_noph36-149-f1.jpg',

'/content/roco/all\_data/train/radiology/images/PMC5754747\_f1000research-6-14507-g0000.jpg',

'/content/roco/all\_data/train/radiology/images/PMC5218828\_nihms839236f3.jpg',

'/content/roco/all\_data/train/radiology/images/PMC3892916\_f1000research-2-1897-g0002.jpg',

'/content/roco/all\_data/train/radiology/images/PMC4345544\_yjbm\_88\_1\_93\_g03.jpg',

'/content/roco/all\_data/train/radiology/images/PMC4345544\_yjbm\_88\_1\_93\_g04.jpg',

'/content/roco/all\_data/train/radiology/images/PMC4728737\_emss-66793-f0004.jpg',

'/content/roco/all\_data/train/radiology/images/PMC4887302\_12471\_2016\_832\_Fig1\_HTML.jpg']

test\_list = ['/content/roco/all\_data/test/radiology/images/PMC5241049\_nihms839240f4.jpg',

'/content/roco/all\_data/test/radiology/images/PMC5357066\_emss-71420-f001.jpg',

'/content/roco/all\_data/test/radiology/images/PMC4544285\_anec0019-0193-f3.jpg']

train\_df = rmv\_empty\_images(train\_df,temp\_list=train\_list)

test\_df = rmv\_empty\_images(test\_df,temp\_list=test\_list)

validation\_df = rmv\_empty\_images(validation\_df,temp\_list=val\_list)

print(train\_df.shape)

print(test\_df.shape)

print(validation\_df.shape)

train\_df = transform\_df(train\_df)

test\_df = transform\_df(test\_df)

validation\_df = transform\_df(validation\_df)

Data Visualizations

uni\_data = train\_df['captions']

uni\_data\_string = ' '.join(uni\_data)

plt.figure(figsize = (20,20))

wc = WordCloud(max\_words = 1000, width=1200, height=600,background\_color="white",stopwords=STOPWORDS).generate(uni\_data\_string)

plt.imshow(wc , interpolation = 'bilinear')

plt.axis('off')

plt.title('Word cloud for Negative reviews',fontsize = 20)

plt.show()